Master Document for IATI Datastore work 2018-2019

Date: 17 december 2018, Revised 30-07-2019,
Author: Siem Vaessen, Zimmerman & Zimmermann

Subject: Living document for the work progress on the IATI Datastore. This document is based on the work as described in the original ToR and the breakdown of the original phases as provided in this document. During the course of the project an open Github based Kanban board is used to monitor progress. Each issue will have a link to the issue on Github for reference.

This document will also make use of colour coding, green having the state 'Done', red having the state 'IATI needs to check', orange meaning 'In Progress' and issues without a colour need to start.' Issues with strikes through have been abandoned, based on the talks for the kick-off meeting of this project after the IATI TAG in Kathmandu. They will not be taken into consideration for production in this scope. See the breakdown of work based on the ToR.

Table of Contents

,

DS Phase 1: Extract, Transform and Load

You can find a status of the tickets for phase 1 @ https://github.com/zimmerman-zimmerman/OIPA/labels/DS%20Phase%201

<table>
<tr><td>A1.1</td><td><b>DS should poll the IATI Registry continuously for updated or new files. Metadata should be stored and made available on the results of each poll (the list of files to be fetched) and each attempted fetch.</b></td></tr>
<tr><td colspan="2">This is done as it is a standard part of OIPA (as part of the default view) named: last_updated_model</td></tr>
<tr><td><b>A1.2</b></td><td><b>DS should access Registry validation logs for each file. (There is no need for DS to perform validation.) Failed files should be ignored. Activities in files with non- critical validation errors should be processed. Only files are fetched which have a current IATI registry validation log.</b></td></tr>
<tr><td colspan="2"><b>This ticket will be parked, once it is clear what Validation Service will be deployed by IATI. It is currently not clear when this new IATI Validation Service will be introduced.</b></td></tr>
<tr><td><b>A1.3</b></td><td><b>Data should be transformed to the latest version of the standard (e.g. Version 2 standardisation of codes, narrative elements, location elements, vocabulary attributes, etc.) [In future, after the next major upgrade, there will be a need for data to be accessed in the two latest versions of the standard.]</b></td></tr>
<tr><td colspan="2"><b>OIPA is no longer supporting Version 1.X of IATI. 30.06.2019 IATI will also deprecate IATI 1.X formally as decided by IATI (IATI TAG 2019). IATI Board needs to make a decision on what makes is into 2.X from 1.X. Needs IATI input to continue.</b></td></tr>
<tr><td><b>A1.4</b></td><td><b>Monetary values should be stored in the original currency and US Dollars (calculated using historical IMF rates for the relevant value dates).</b></td></tr>
<tr><td colspan="2">This is part of OIPA. It stores the original currency and make a conversion to U$, EUR, SDR and others. However, we need to double check the mechanism on how the conversion is done and how we can get to the historical IMF Rates. Also: how far back do we need to go in time? First available project activity in IATI perhaps as a baseline? And what if another project becomes ‘earlier’?</td></tr>
</table>

<table>
<tr><td colspan="2">We use IMF for currency conversion inside of the following Python script as part of OIPA: http://bit.ly/2siLKHd</td></tr>
<tr><td><strong>A1.5</strong></td><td><strong>A copy of the most recent original xml should be stored for each activity.</strong></td></tr>
<tr><td colspan="2">OIPA does ont store the latests activity files for each IATI publisher that is registered on the IATI Registry but it depends on the publisher itself,</td></tr>
<tr><td><strong>A1.6</strong></td><td><strong>Processing and error handling during ETL needs to handle known complexities, including the reporting of the same activity twice, temporarily missing activities, etc</strong></td></tr>
<tr><td colspan="2">Currently we make use of log files, so everything is logged. If required we could integrate Sentry for this and have a more granular view of those logfiles.</td></tr>
<tr><td><strong>A1.7</strong></td><td><strong>All logic employed in transformation processes must be clearly and transparently documented for both developers and end users.</strong></td></tr>
<tr><td colspan="2">We have provided a list of ETL naming ‘conventions’ here: https://docs.google.com/spreadsheets/d/14SNYG5n4e582nbVijBzlIUitQBA8iW0_w2N6WMmrFIE/edit?usp=sharing it provides a list of actual IATI fields name according how they are named in the standard and how they are used in the OIPA Datamodel for version 2.0.1 - 2.0.3</td></tr>
<tr><td>~~A1.8~~</td><td>~~Metadata should be maintained that records all validation errors (via Registry) and transformation processes for each activity.~~</td></tr>
<tr><td><strong>A1.9</strong></td><td><strong>Data should be stored in a manner that maximises speed and flexibility of query and access.</strong></td></tr>
<tr><td colspan="2">OIPA supports this.</td></tr>
</table>

DS Phase 1: Filters Querying should be possible using the following filters

B1.1 Last updated date and time

This has been added to the model.

B1.2 All elements of the standard

This work has been split into different issues to compare the IATI XML standard and version to the relevant data models inside of OIPA. Our approach is to have a comparison for both the activity and organisation file and review them for validation. We will then move to make sure the OIPA data model has a 100% coverage. Each element of the standard in the data model will / has a test for it. Not each field inside of OIPA has an API call available, we will review which elements need to be queryable.

We have split this task into 3 subtasks, covering 2.0.3, 2.0.2 and 2.01.

B1.2.1 All elements of IATI V2.0.3

We have been working on providing a complete list of all the IATI elements and its types and made a comparison to the OIPA data model. Based on that, we can conclude that the OIPA data model for this version is near complete. We will make sure to add the missing element and some types to the data model so it will be complete. The document with the comparison can be found here, it will have multiple tabs for each version of IATI.

We identified some elements not parsed and are nearing fil integration. See https://docs.google.com/document/d/1op2qAf-E53MiKWlva6P8OIiZqdREvd_0jFy-1OcXhQA/edit for update. Team tells us out of the 14 fields not covered.

B1.2.2 All elements of IATI V2.0.2

We have been working on providing a complete list of all the IATI elements and its types and made a comparison to the OIPA data model. Based on that, we can conclude that the OIPA data model for this version is near complete. We will make sure to add the missing element and some types to the data model so it will be complete. The document with the comparison can be found here, it will have multiple tabs for each version of IATI.

The 2.0.2 data is parsed using the 2.0.3 parser. No unique parser is used.

B1.2.3 All elements of IATI V2.0.1

We have been working on providing a complete list of all the IATI elements and its types and made a comparison to the OIPA data model. Based on that, we can conclude that the OIPA data model for this version is near complete. We will make sure to add the missing element and some types to the data model so it will be complete. The document with the comparison can be found here, it will have multiple tabs for each version of IATI.

The 2.0.1 data is parsed using the 2.0.2 parser (equals) 2.0.3 parser. No unique parser is used.

B1.3 Free text search across title, description or entire activity

OIPA supports this. See https://yoda.oipa.nl/api/activities/ for search coverage. See TT request.

API request may include q parameter. This parameter controls text search and contains expected value. By default, searching is performed on:

- iati_identifier the IATI identifier
- title narratives
- description narratives
- recipient_country recipient country code and name
- recipient_region recipient region code and name
- reporting_org ref and narratives
- sector sector code and name
- document_link url, category and title narratives
- participating_org ref and narratives

To search on subset of these fields the q_fields parameter can be used, like so: q_fields=iati_identifier,title,description By default, search only return results if the hit resembles a full word. This can be altered through the q_lookup parameter. Options for this parameter are:

- exact (default): Only return results when the query hit is a full word.
- startswith: Also returns results when the word stars with the query.

B1.4 Pre-processed aggregations and calculations fields, including: US Dollar monetary values; CRS Sectors; Years (from dates)

We will first update the documentation to see where OIPA stands with this aggregation.

Documented here: https://github.com/zimmerman-zimmerman/OIPA/issues/856

Current aggregations:

- budget_value & budget_currency
- disbursement_value & disbursement_currency
- incoming_funds_value & incoming_funds_currency
- commitment_value & commitment_currency
- expenditure_value & expenditure_currency
- interest_payment_value & interest_payment_currency
- loan_repayment_value & loan_repayment_currency
- reimbursement_value & reimbursement_currency
- purchase_of_equity_value & purchase_of_equity_currency
- sale_of_equity_value & sale_of_equity_currency
- credit_guarantee_value & credit_guarantee_currency
- incoming_commitment_value & incoming_commitment_currency

Data model:
https://github.com/zimmerman-zimmerman/OIPA/blob/develop/OIPA/iati/models.py#L454

Process aggregation:
https://github.com/zimmerman-zimmerman/OIPA/blob/develop/OIPA/iati/activity_aggregation_calculation.py#L72

US Dollar monetary values:
The process aggregation is working only to the specific currency which has related to the current record. For example: if the current record is EUR then the aggregation is EUR, currently is no data conversion to USD.

CRS Sectors:
The aggregation process doesn't cover the sector data.

Years (from dates):
The process aggregation to all data related to the activity.

Example on how to use this function

DS Phase 1: Developer Output

C1.1	**Developer queries should result in the appropriately filtered delivery of entire standardised activities in XML format**
Moving this to Phase 2 working schedule. **Example on how to use this function**	
C1.2	**Developer queries should result in the appropriately filtered delivery of entire standardised activities in JSON format.**
Example on how to use this function	
C1.3	**Developer queries should result in the appropriately filtered delivery of entire standardised activities in original (untransformed) XML.**
Example on how to use this function	
C1.4	**The standardised output should follow the standard IATI XML as close as possible.**
OIPA has been following the IATI standard as leading for its data model. The core of OIPA does not have any room for data model modifications that are non native to IATI.	

DS Phase 1: API

D1.1	**Any products making use of the DS, including the DS query interface itself, may only communicate with the datastore via the API.**

OIPA has been built as a RESTful service. It can be openly accessed via its API and a separate administrator Django based administrative area.	
D1.2	**The API must be RESTful, must be versioned, must use HTTPS**
We make use of Letsencrypt SSL certificate generation and propose to include those on the Hosting environments as well unless IATI required to sign-off on another type of certificate supplier.	
~~D1.3~~	~~Results should be deliverable in full.~~
D1.4	**Results should be deliverable paginated,**
The OIPA API provide the option to paginate results.	

END OF PHASE 1 WORK ITEMS

DS Phase 2: CSV/XLSX Serialisations

You can find a status of the tickets for phase 2 @ https://github.com/zimmerman-zimmerman/OIPA/labels/DS%20Phase%202

E2.1	**CSV and XLSX outputs should be available in the following serialisations - defined by the unit represented on each row:**
E2.1.1	Organisation
E2.1.2	Activity
E2.1.3	Transaction
E2.1.4	Activity Budget
E2.1.5	Organisation Budget
E2.1.6	Result
E2.1.7	Location
E2.1.8	Document links
E2.2	XLSX output should allow for multiple serialisations to selected to build a multi- tab workbook.
E2.3	Transaction and Budget serialisations should also allow for data rows exploded by recipient country/region and/or sector where multiple values for country/region and/or sector exist
E2.4	Each serialisation should be built with a default set of columns [tbc] but users should have the option to make their own selection of columns to display.

DS Phase 2: Analyst Interface

F2.1	A functional user interface should allow analysts to build ~~complex~~ queries on both activities and organisations that:
F2.1.1	Result in csv or xlsx file downloads
F2.1.2	Utilise the filters outlined above
F2.1.3	Choose from a number of serialisations outlined above
F2.1.4	Select the columns to display within each serialisation

DS Phase 2: Publisher Interface

~~G2.1~~	~~Provide a report that allows publishers to check the status of their data in both the Registry and Data store~~

END OF PHASE 2 WORK ITEMS

DS Phase 3: Hosting & maintenance

H3.1	Phase 3: Maintenance/Hosting services - The bidder must undertake to host the DataStore for a period until 31 December 2019.
H3.1	Phase 3: The bidder should confirm availability to provide hosting/maintenance services beyond 31 December 2019 for a period of one year, providing the estimated annual cost

END OF PHASE 3 WORK ITEMS